SOLUTION BRIEF

Server Performance

Providing On-Premise Object Storage for MariaDB

Decoupling the storage media from database servers means storage can be scaled out and used more efficiently. Using RocksDB-Cloud and MinIO, you can use shared object storage for databases running in your data center

Executive Summary

Faced with huge volumes of data, businesses need to find cost-effective ways to manage it. In the past, databases have run on servers with their own attached storage. That makes it difficult to scale, and to optimize resource use across the data center. While hyperscale cloud providers offer databases based on shared storage, some companies may be reluctant to put their data in the cloud. Proprietary, on-premise solutions are available, but they may result in vendor lock-in and can incur significant licensing costs.

Now, there is a way that organizations can keep their data on premise and also optimize their data center resources using shared storage for the MariaDB database. The solution uses Rockset's RocksDB-Cloud to add shared object storage to MariaDB, together with the MinIO open-source server software for object storage on premise. Because the software components are open source, there is no vendor lock-in and there are no licensing costs. There is no speed penalty, either: testing shows that the solution performs as fast as the shared-nothing MariaDB configuration[1].

Using the 2nd generation Intel® Xeon® Gold processor with Intel® Optane™ DC persistent memory enables the block cache to be turned off. That results in significant cost savings by reducing DRAM capacity.

Solution Benefits

- **Scalable storage.** Separating the compute and storage resources enables them to be scaled independently. The persistent local storage used for the full database image can be replaced with object storage.
- **Higher utilization**. Using a shared storage pool enables higher average utilization rates than would be expected with systems where the storage hardware is coupled to the database engine.
- **On-premise deployment.** While the solution supports cloud deployments, it can also be used on premise using MinIO for object storage locally. For companies that cannot or prefer not to put their data into the cloud, this gives them the benefit of shared object storage within their own data center.
- **Open source.** Companies can avoid the expense and lock-in of proprietary software by using open source for the database and other components of the solution.

Authors

David Cohen
Storage Solutions CTO & Sr PE
DCG/NVMS

Steve Shaw
Application Engineer
CEE ESEE ODA

Mikhail Sinyavin
Application Engineer
CEE CSE VCE

Rino Cavallucci
Software enabling Project Mgr.
IAGS DRS LEAP

Business Challenge: Improving Database Efficiency with Shared Storage

Businesses today need a strategy to handle large and growing data volumes efficiently. Traditionally, databases have run on servers with their own, direct-attached storage devices. This limits the size of the database to the storage capacity of the server where it runs. Secondly, it is difficult to optimize the storage resources as a whole since spare capacity in one storage device cannot be used by a database on a different server. As a result, there may be wasted capacity. Finally, it is harder to scale solutions where the data and the database engine are tightly coupled together.

To increase efficiency, a shared storage model can now be used instead. This increases the utilization rate of the storage media, by enabling the storage pool to be shared across database instances, and enables resources to be scaled in line with demand. Hyperscale cloud providers are making databases available in the cloud using shared storage, but some businesses hesitate to put their most sensitive data into the cloud. On-premise solutions using scale-out storage are available commercially, but businesses may prefer to avoid proprietary solutions with the associated expense and vendor lock-in.

To increase database performance and retain full oversight of the data, businesses need an on-premise, shared-storage database solution, preferably based on open-source software.

Solution Value: Cloudifying MariaDB

The solution described in this paper enables cloud-native object storage to be used for data stored in a MariaDB database. The scale-out storage is enabled without incurring any performance penalty. The enabler is configuring MariaDB to use a version of the MyRocks storage engine linked to Rockset's RocksDB-Cloud library.

The solution, in common with other MariaDB/MyRocks deployments, stores the hottest data in memory. DRAM is expensive, though, so the solution uses Intel Optane DC persistent memory as low latency, local storage, at a lower cost per bit stored[2]. The solution shows that the block cache can be turned off when using persistent memory[1], reducing the cost of DRAM.

The solution is based on open-source software, so it can be used without concerns about vendor lock-in and without the expense of software licenses.

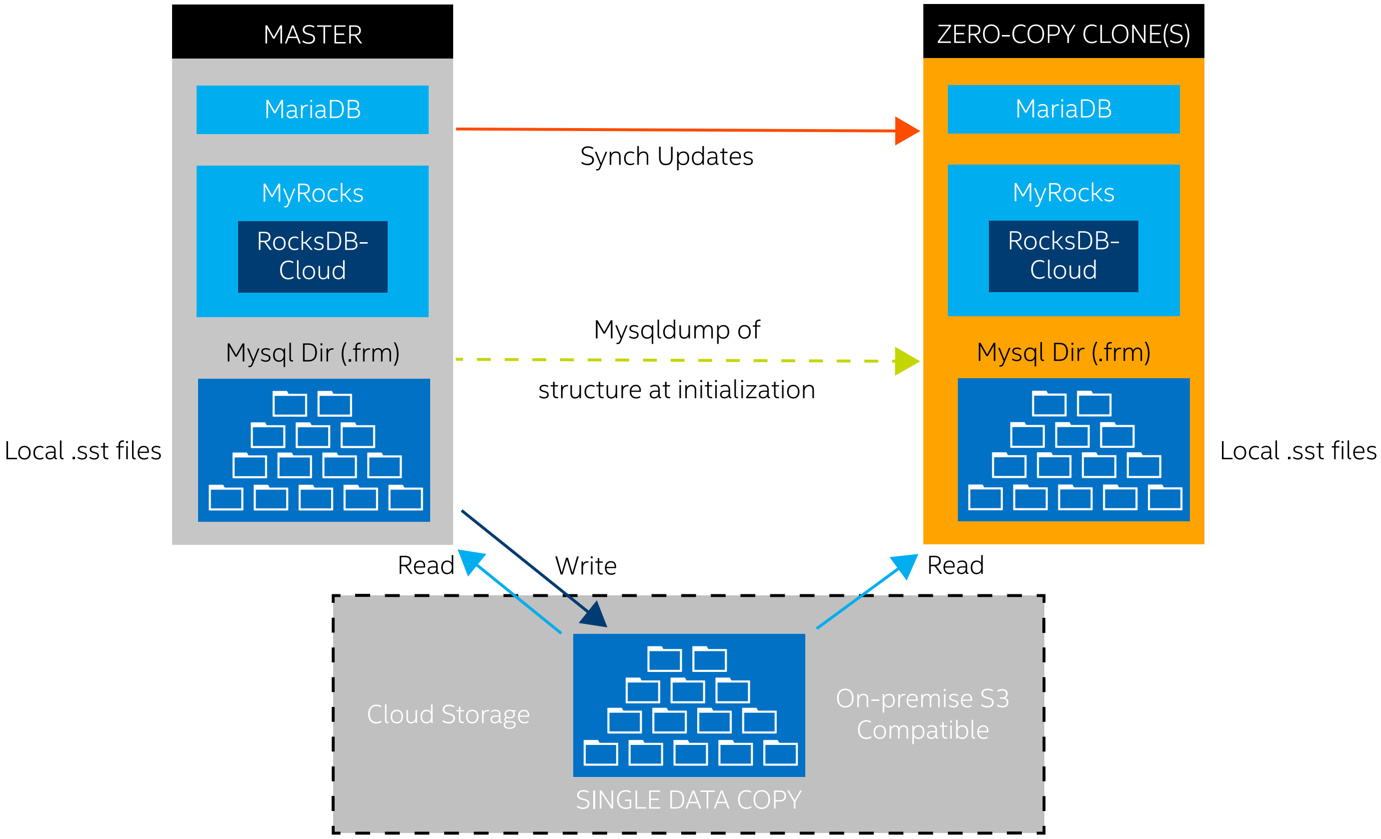

Figure 1. The solution architecture shows how zero-copy clones can be used to scale up the database's compute capacity, while cloud storage, or on-premise object storage using MinIO, can be used to scale up storage.

Solution Architecture: On-Premise Cloud Native Database

When deployed with MyRocks storage engine, MariaDB is a caching database, where the full image of the database is stored on lower-cost, persistent, local storage while the active state is stored in memory. Changes are appended to a write-ahead log (WAL) before being applied to MemTables, which are also known as in-memory write-buffers. When a MemTable becomes full it is marked as immutable. Subsequently, its contents are flushed to a sorted string table (SST) file that resides on persistent, local storage and the WAL entries associated with the MemTable are marked for deletion.

The new clustered database design uses separate compute and storage nodes, so compute and storage can be scaled independently. To enable the persistent state to be stored using object storage protocols, the RocksDB-Cloud library created by Rockset can be used. RocksDB-Cloud is a wrapper around the RocksDB library that enables the MyRocks storage engine to use object storage. The object storage can be in the cloud, but to provide a scalable, on-premise solution, we use MinIO, open-source server software compatible with Amazon S3 APIs.

Backed by Intel SSDs, a Minio Bucket holds the full database image. The combination of MinIO's distributed erasure codes and Intel's high density SSDs greatly reduce the data center footprint compared to using hard drives. Intel Optane DC persistent memory is used for the active state (or cache), to accelerate reads and writes in the compute nodes. Persistent memory is a breakthrough in the memory hierarchy, offering near-DRAM speeds at more affordable high capacities.
The servers are based on 2nd generation Intel Xeon Gold processors that offer workload-optimized performance and advanced reliability in support of demanding data center and analytics applications.

Zero copy clones can be used to scale up database reads (see Figure 1). The clone accesses the object storage directly for older data, and stays synchronized with the master using Binlog Semi-Sync replication. MariaDB Maxscale can be used to balance the read and write load across the cluster.

The solution enables sharding to be used to protect the data across fault domains, so the amount of redundancy for data recovery purposes can be cut from 3x (master copy plus two replicas) to 1.5x[3].

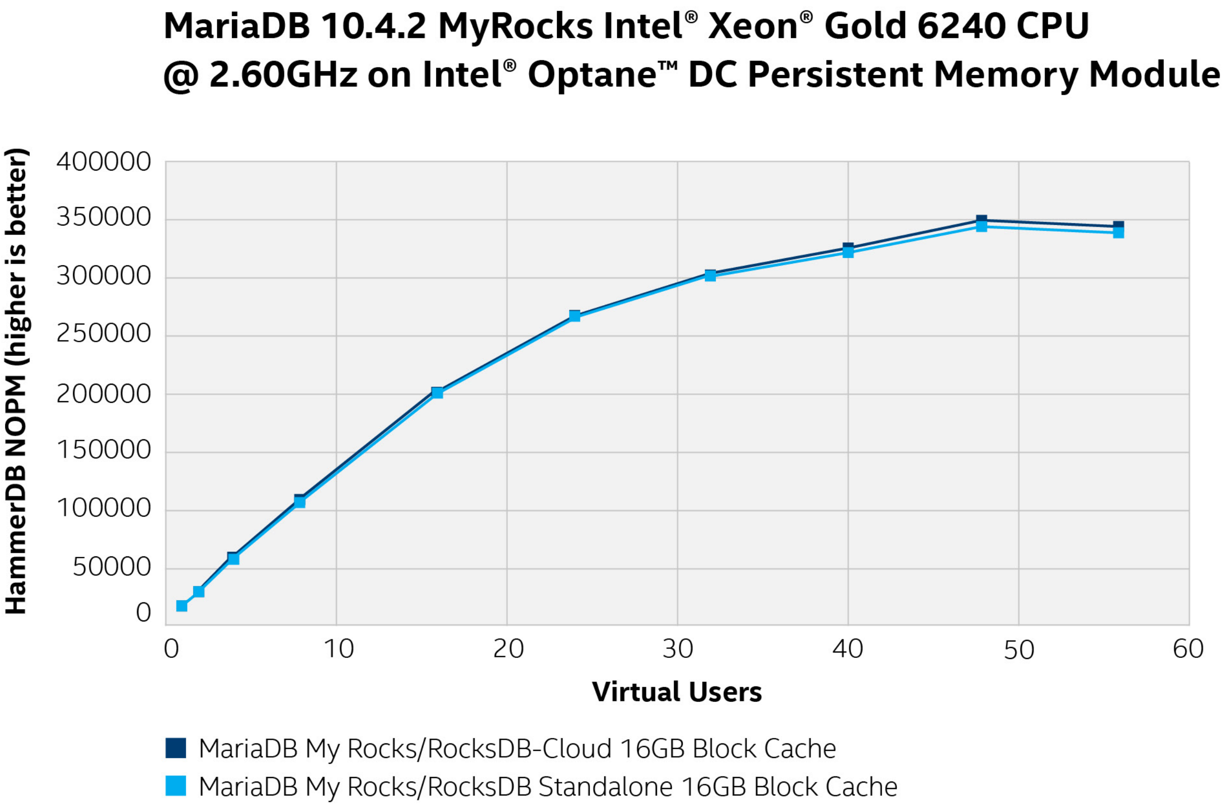

Figure 2. Performance test results (an average of two runs) showing new orders per minute (NOPM). Higher is better. These results show that RocksDB-Cloud configurations match the performance of standalone clusters.

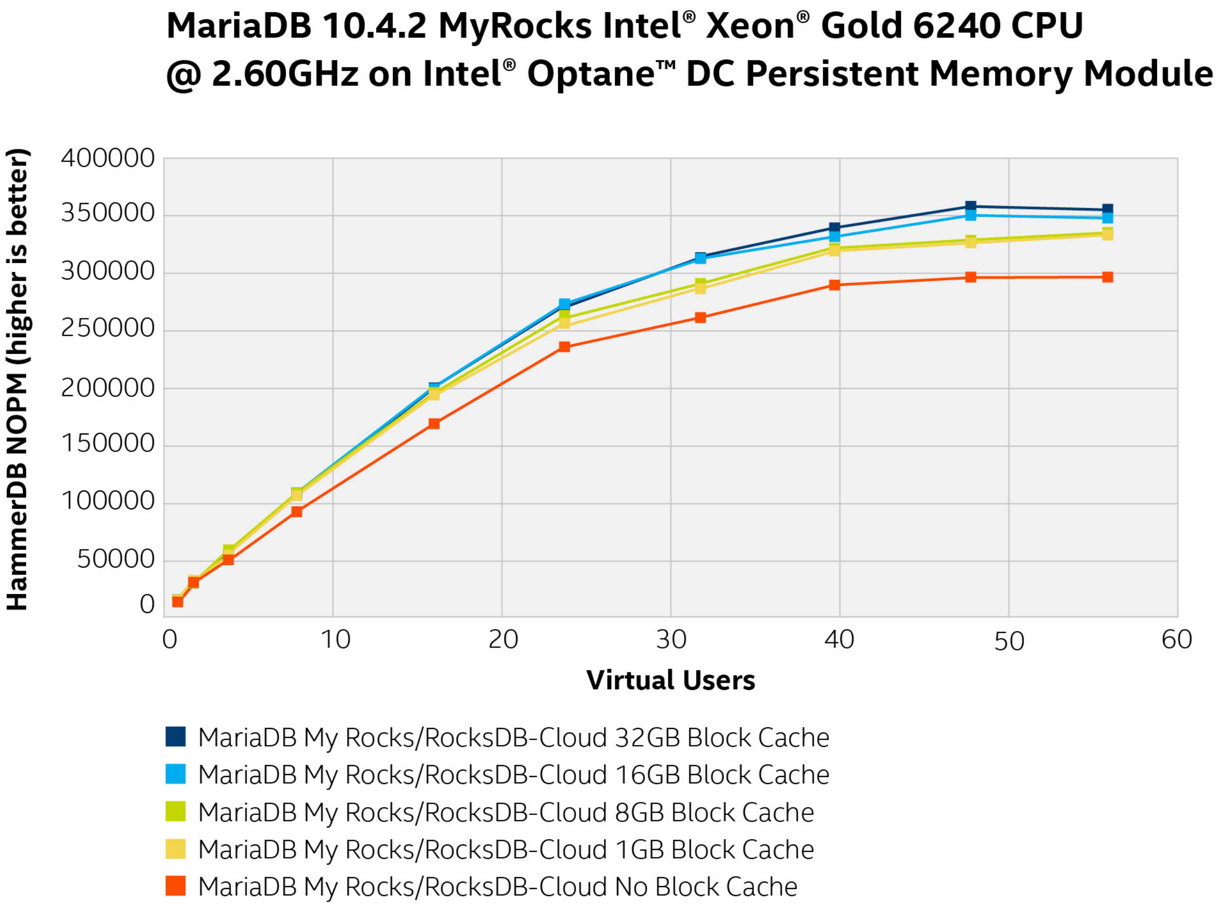

Figure 3. Testing the minimum effective block cache size using Intel Optane DC persistent memory[1]. Graph shows new orders per minute (higher is better). There is a small reduction in performance with each block cache reduction. This graph shows that with the data in Intel Optane DC persistent memory, we can turn the block cache off while maintaining good performance, making a significant saving in DRAM.

Performance Testing

Performance testing was carried out to confirm that the shared-storage architecture was as fast as the shared-nothing architecture MariaDB uses by default. Testing was carried out using HammerDB, an open-source load testing tool that implements a CPU intensive simulated workload.

Testing showed that RocksDB-Cloud configurations matched the performance of standalone cluster tests (see Figure 2)[1]. There is no overhead in copying the data to object storage.

Tests were also carried out to measure the impact of Intel Optane DC persistent memory. The block cache is used for writing the SST tables to the persistent storage. A cache size of 1GB, 8GB, 16GB and 32GB was tested, as well as disabling the block cache altogether. Figure 3 shows the results, which were achieved without any code changes. They show a small reduction in performance at each step from 32GB to no block cache at all[1]. With the data in Intel Optane DC persistent memory, we can turn the block cache off, making a significant saving in DRAM. This is especially important in cloud environments.

Conclusion

It is possible to use MariaDB with shared object storage, which can be in the cloud or on-premise. The test results show that the architecture is as fast as the default shared-nothing configuration of MariaDB. Using Intel Optane DC persistent memory, it is possible to turn the block cache off, greatly reducing the amount of relatively expensive DRAM that is required. The solution is based on open-source software, providing vendor neutrality and eliminating software license costs.

Learn More

- **2nd Generation Intel Xeon Gold processor**
- **Intel Optane DC persistent memory**
- **RocksDB-Cloud**

Find the solution that is right for your organization. Contact your Intel representative or visit **intel.com**.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks

[1] Configurations: Testing carried out by Intel September 2019. Intel Xeon Gold 6240 processor @ 2.60GHz; 2 sockets; six 128GiB Intel Optane DC persistent memory DIMMs per socket; six 32GiB DRAM DIMMs per socket; cache size: 25344 KB. Testing carried out using HammerDB with 400 warehouses. Database platform: MariaDB 10.4.2, RocksDB-Cloud version 5.17.2, MyRocks storage engine version 1.0. Operating system: Red Hat Enterprise Linux Server release 7.6 (Maipo) - 3.10.0-957.21.3.el7.x86_64 #1 SMP Fri Jun 14 02:54:29 EDT 2019 x86_64 x86_64 x86_64 GNU/Linux

[2] Intel Optane DC persistent memory pricing & DRAM pricing referenced in TCO calculations is provided for guidance and planning purposes only and does not constitute a final offer. Pricing guidance is subject to change and may revise up or down based on market dynamics. Please contact your OEM/distributor for actual pricing.

[3] This is a function of using Minio's distributed erasure codes compared to (n)-way replication. Erasure codes have a redundancy factor of k where k is the number of parity blocks. In an 8+4 EC stripe k=4 so the redundancy factor is 1.5. In contrast, using (n)-way replication the redundancy factor is always 2 or greater. 3 is typical.

Intel does not control or audit third-party data. You should review this content, consult other sources, and confirm whether referenced data are accurate.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. **No product or component can be absolutely secure**. Check with your system manufacturer or retailer or learn more at intel.com.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Intel technologies may require enabled hardware, specific software, or services activation. Check with your system manufacturer or retailer.